

# DATA-INTENSIVE RESEARCH IN ECOLOGY: COMBINING LARGE DATASETS, BEST PRACTICES, AND COMPUTATIONAL TRAINING

Kristina Riemer  
University of Florida

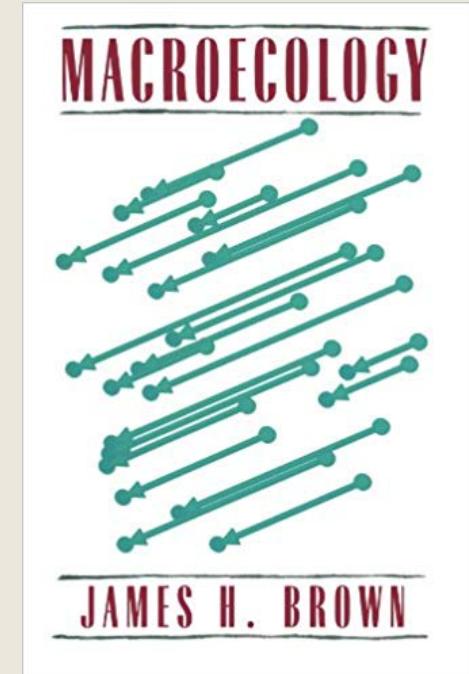
# The start of my computing journey



# The start of my computing journey



# The start of my computing journey



# What determines organism size?

Opinion

Cell  
PRESS

## Declining body size: a third universal response to warming?

Janet L. Gardner<sup>1</sup>, Anne Peters<sup>2,3</sup>, Michael R. Kearney<sup>4</sup>,  
Leo Joseph<sup>5</sup> and Robert Heinsohn<sup>1</sup>

<sup>1</sup>Fenner School of Environment and Society, Australian National University, Canberra, ACT 0200, Australia

<sup>2</sup>Behavioral Ecology of Sexual Signals Group, Max Planck Institute for Ornithology, Vogelwarte Radolfzell,  
78315 Radolfzell, Germany

<sup>3</sup>School of Biological Sciences, Monash University, VIC 3168, Australia

<sup>4</sup>Department of Zoology, The University of Melbourne, VIC 3010, Australia

<sup>5</sup>Australian National Wildlife Collection, CSIRO Ecosystem Sciences, GPO Box 284, Canberra, ACT 2601, Australia

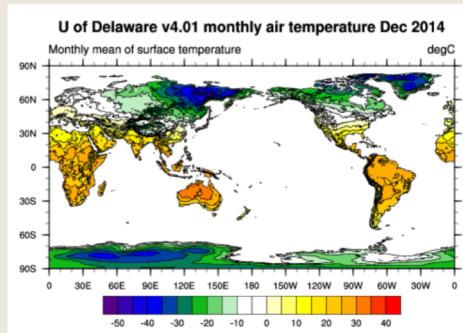
- Smaller when warmer
- One of few "rules" in ecology
- Datasets with small sample sizes

Data-intensive approach to size  
response to temperature



15 million museum records

# Data-intensive approach to size response to temperature

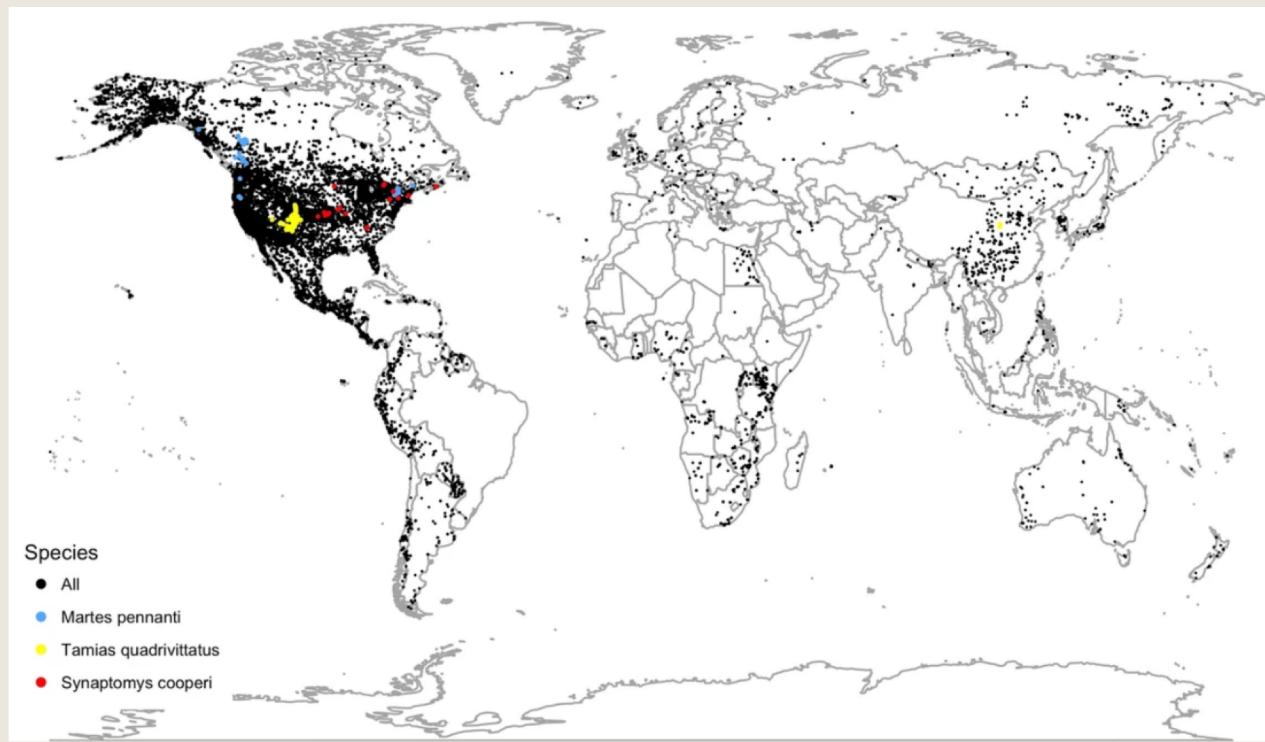


15 million museum records

+

Raster of >250,000 grids

# Data-intensive approach to size response to temperature

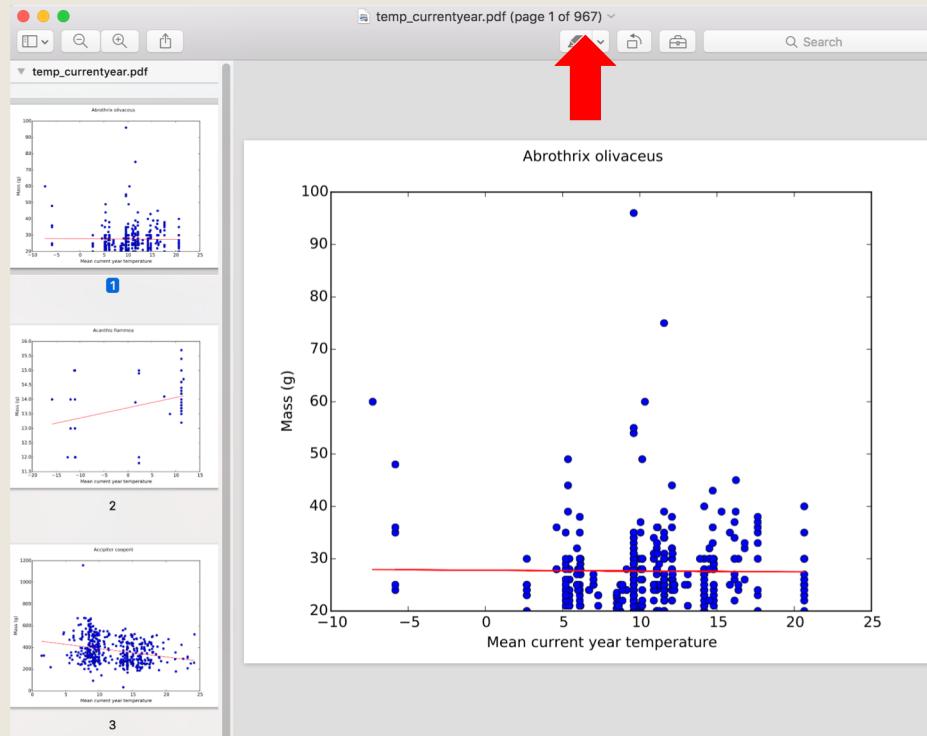


# New challenges required new skills

```
#-----FUNCTIONS-----
download_VN = function(raw_file_path){
  # Download organismal data (Vertnet)
  #
  # Args:
  #   raw_file_path: file path to raw data
  #
  # Returns:
  #   Single csv of four desired class-level datasets with only necessary columns
  #   and only rows with mass values
  if(!file.exists(raw_file_path)){
    rdataretriever::install("vertnet-amphibians", "sqlite", db_file = "data/all_vertnet.db")
    rdataretriever::install("vertnet-birds", "sqlite", db_file = "data/all_vertnet.db")
    rdataretriever::install("vertnet-mammals", "sqlite", db_file = "data/all_vertnet.db")
    rdataretriever::install("vertnet-reptiles", "sqlite", db_file = "data/all_vertnet.db")
    database = src_sqlite("data/all_vertnet.db")
    amphibians_query = "SELECT scientificname, class, ordered, family, year, decimallongitude
birds_query = "SELECT scientificname, class, ordered, family, year, decimallongitude, dec
mammals_query = "SELECT scientificname, class, ordered, family, year, decimallongitude, d
reptiles_query = "SELECT scientificname, class, ordered, family, year, decimallongitude,
query = paste(amphibians_query, "UNION ALL", birds_query, "UNION ALL", mammals_query, "UN
subset_data = tbl(database, sql(query))
write.csv(subset_data, file = raw_file_path)
  }
}
```

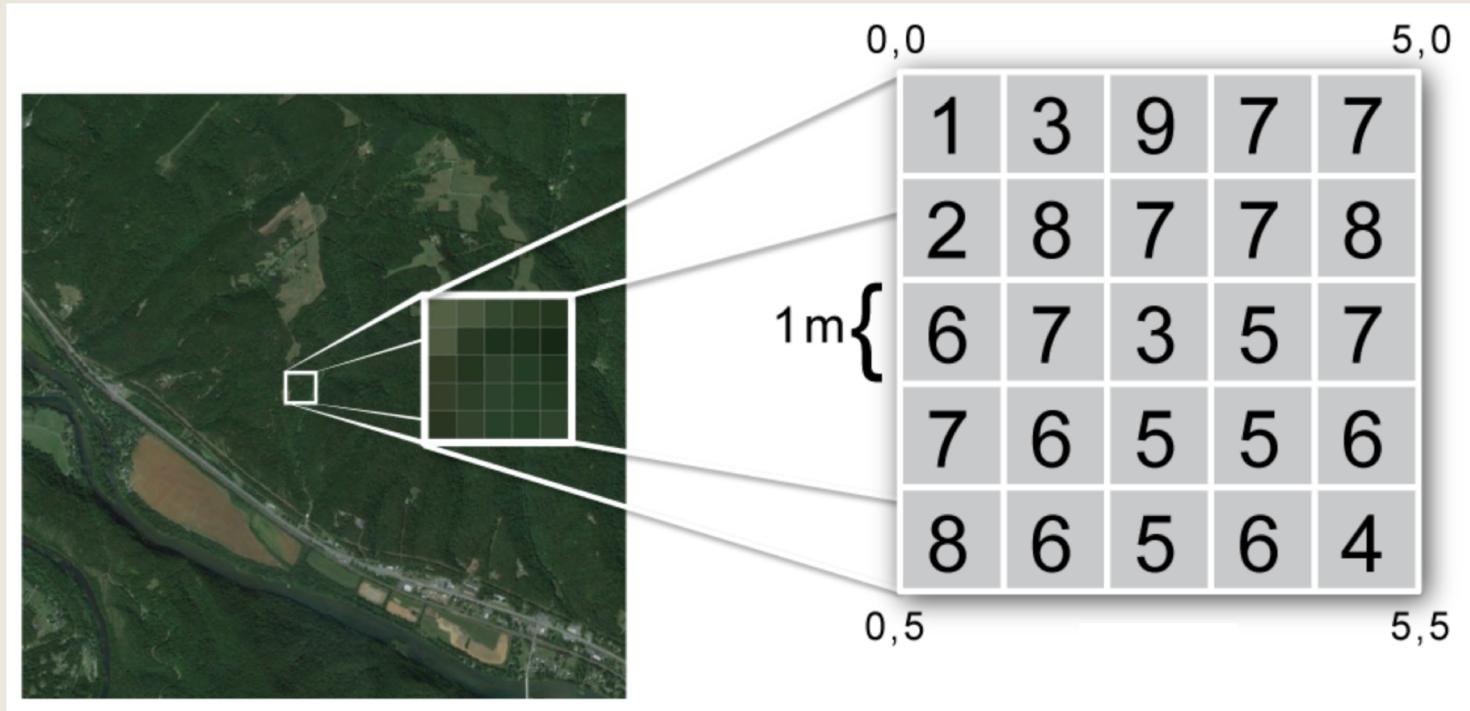
1. Very large amounts of data

# New challenges required new skills



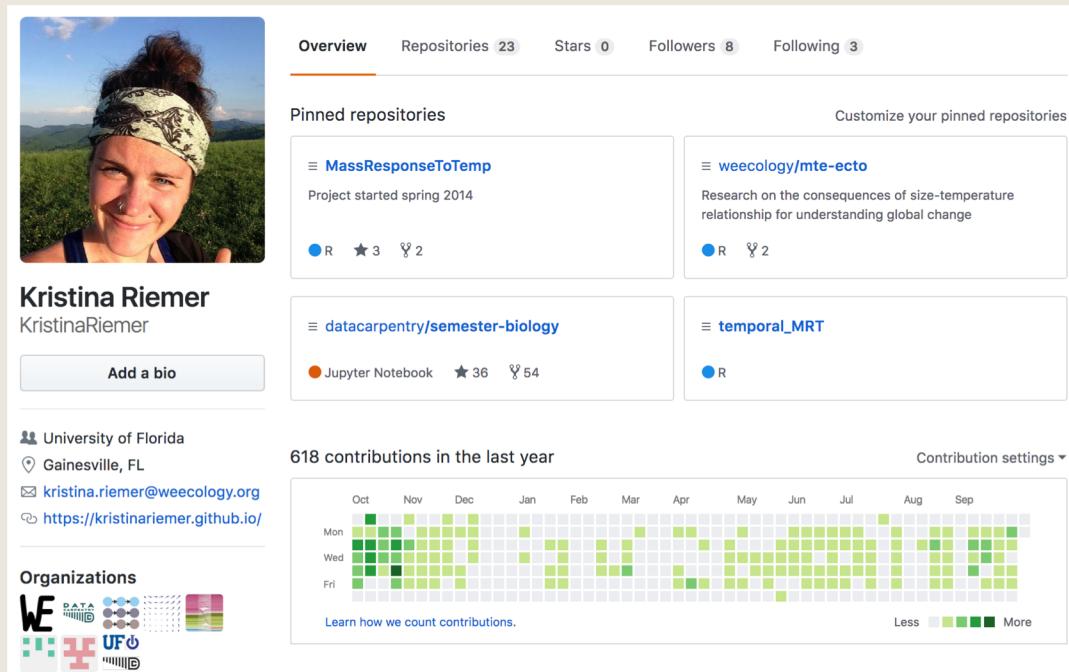
## 2. Repeated tasks on lots of data

# New challenges required new skills



## 3. Spatial data

# New challenges required new skills



Kristina Riemer's GitHub profile page. The top navigation bar shows "Overview" (selected), "Repositories 23", "Stars 0", "Followers 8", and "Following 3". Below the navigation, there are four pinned repositories:

- MassResponseToTemp**: Project started spring 2014. Language: R, Stars: 3, Forks: 2.
- weecology/mte-ecto**: Research on the consequences of size-temperature relationship for understanding global change. Language: R, Forks: 2.
- datacarpentry/semester-biology**: Jupyter Notebook, Stars: 36, Forks: 54.
- temporal\_MRT**: Language: R.

The main content area shows "618 contributions in the last year" with a heatmap visualization. The heatmap grid shows contributions by day of the week (Mon, Tue, Wed, Thu, Fri) and month (Oct, Nov, Dec, Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep). A legend indicates that lighter shades represent "Less" contributions and darker shades represent "More" contributions. Below the heatmap, it says "Learn how we count contributions." and "Contribution settings ▾".

## 4. Tracking code

# Success!

KristinaRiemer / MassResponseToTemp

Code Issues 1 Pull requests 0 Projects 0 Wiki Insights Settings

Project started spring 2014 Edit

Manage topics

414 commits 4 branches 1 release 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

KristinaRiemer Update main figures for proofs Latest commit 7bff70d on Dec 7, 2017

File	Description	Time Ago
Analysis_VN_CY.py	Add float32 conversion to analysis scripts	a year ago
Analysis_VN_TL.py	Add float32 conversion to analysis scripts	a year ago
Cleaning_VN.R	Add in missing column	2 years ago
MRT_manuscript.doc	Add post-acceptance edits	11 months ago
README.md	Add DOI to README	11 months ago
Visualization_VN_CY.R	Update main figures for proofs	10 months ago
Visualization_VN_CY_supp.R	Move z score figure to supplement figs script	11 months ago
environment.yml	Use conda for all Python package installation	a year ago
install-packages.R	Create conda env and R script for automated package install	2 years ago
repro_run.sh	Add ability to run full analysis using a bash argument	a year ago

# Success!

KristinaRiemer / MassResponseToTemp

Code Issues 1 Pull requests 0 Projects 0 Wiki Insights Settings

Project started spring 2014 Edit

Manage topics

414 commits 4 branches 1 release 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

KristinaRiemer Update main figures for proofs Latest commit 7bff70d on Dec 7, 2017

File	Description	Time Ago
Analysis_VN_CY.py	Add float32 conversion to analysis scripts	a year ago
Analysis_VN_TL.py	Add float32 conversion to analysis scripts	a year ago
Cleaning_VN.R	Add in missing column	2 years ago
MRT_manuscript.doc	Add post-acceptance edits	11 months ago
README.md	Add DOI to README	11 months ago
Visualization_VN_CY.R	Update main figures for proofs	10 months ago
Visualization_VN_CY_supp.R	Move z score figure to supplement figs script	11 months ago
environment.yml	Use conda for all Python package installation	a year ago
install-packages.R	Create conda env and R script for automated package install	2 years ago
repro_run.sh	Add ability to run full analysis using a bash argument	a year ago

# Success!

KristinaRiemer / MassResponseToTemp

Code Issues 1 Pull requests 0 Projects 0 Wiki Insights Settings

Project started spring 2014 Edit

Manage topics

414 commits 4 branches 1 release 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

KristinaRiemer Update main figures for proofs Latest commit 7bff70d on Dec 7, 2017

Analysis_VN_CY.py	Add float32 conversion to analysis scripts	a year ago
Analysis_VN_TL.py	Add float32 conversion to analysis scripts	a year ago
Cleaning_VN.R	Add in missing column	2 years ago
MRT_manuscript.doc	Add post-acceptance edits	11 months ago
README.md	Add DOI to README	11 months ago
Visualization_VN_CY.R	Update main figures for proofs	10 months ago
Visualization_VN_CY_supp.R	Move z score figure to supplement figs script	11 months ago
environment.yml	Use conda for all Python package installation	a year ago
install-packages.R	Create conda env and R script for automated package install	2 years ago
repro_run.sh	Add ability to run full analysis using a bash argument	a year ago

# Success!

KristinaRiemer / MassResponseToTemp

Code Issues 1 Pull requests 0 Projects 0 Wiki Insights Settings

Project started spring 2014 Edit

Manage topics

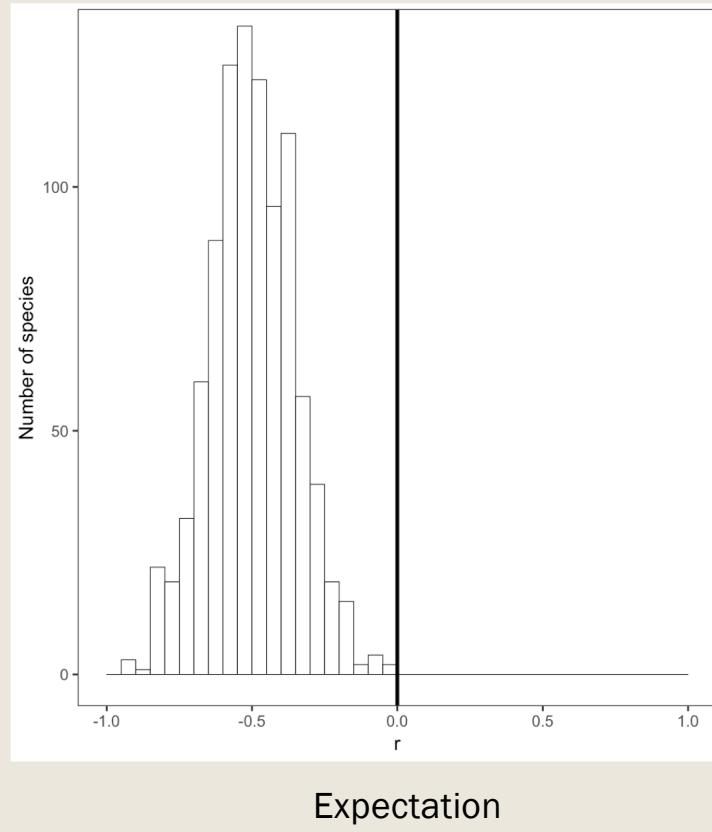
414 commits 4 branches 1 release 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

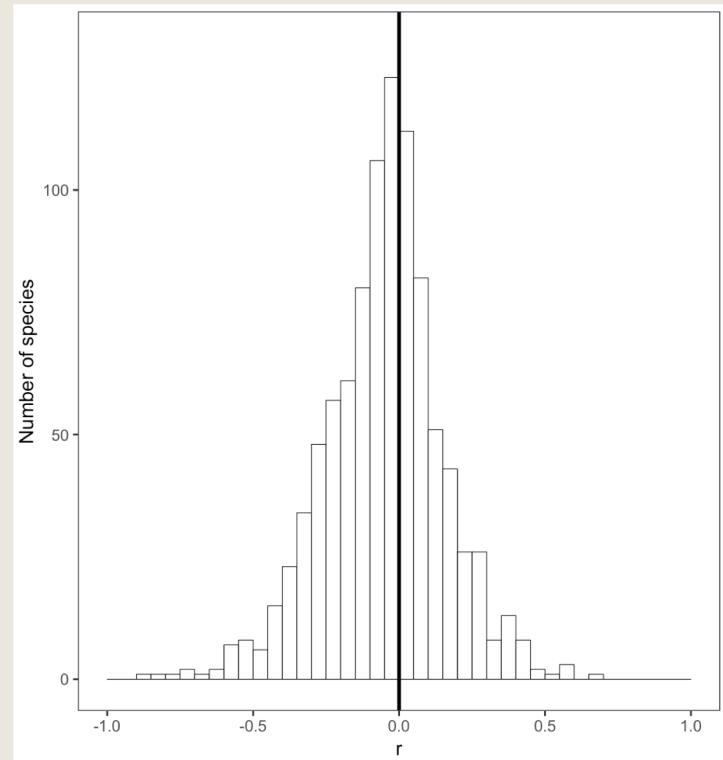
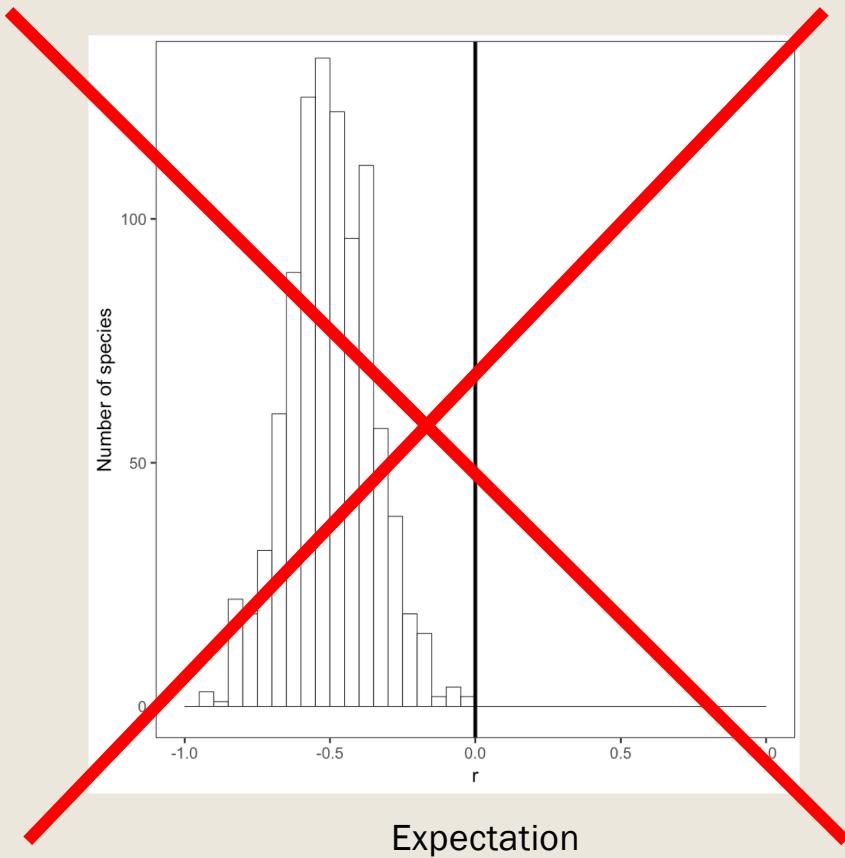
KristinaRiemer Update main figures for proofs Latest commit 7bff70d on Dec 7, 2017

File	Description	Time Ago
Analysis_VN_CY.py	Add float32 conversion to analysis scripts	a year ago
Analysis_VN_TL.py	Add float32 conversion to analysis scripts	a year ago
Cleaning_VN.R	Add in missing column	2 years ago
MRT_manuscript.doc	Add post-acceptance edits	11 months ago
<b>README.md</b>	Add DOI to README	11 months ago
Visualization_VN_CY.R	Update main figures for proofs	10 months ago
Visualization_VN_CY_supp.R	Move z score figure to supplement figs script	11 months ago
environment.yml	Use conda for all Python package installation	a year ago
install-packages.R	Create conda env and R script for automated package install	2 years ago
repro_run.sh	Add ability to run full analysis using a bash argument	a year ago

# Lack of evidence for ecology “law”



# Lack of evidence for ecology “law”

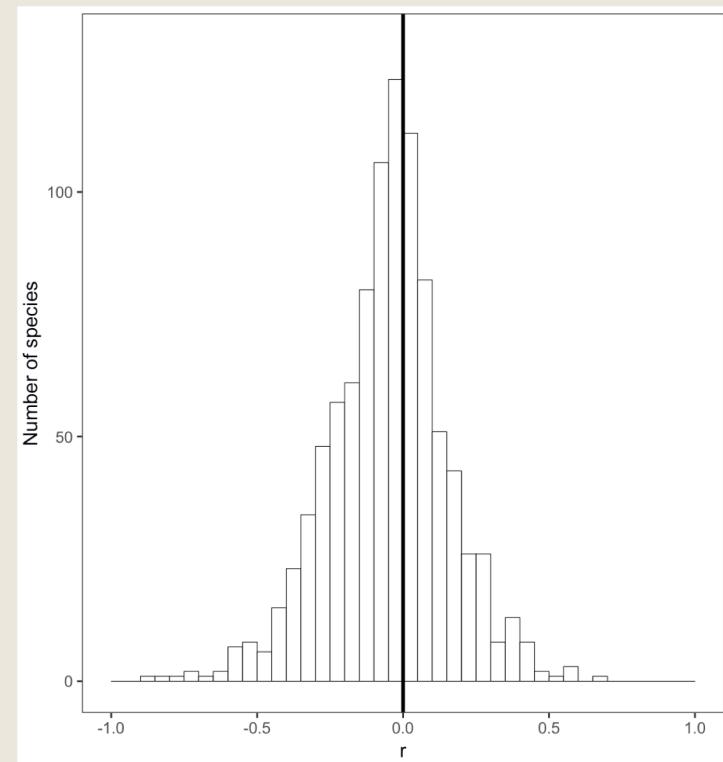


Expectation

Results from data-driven approach

# Lack of evidence for ecology “law”

“These results suggest that Bergmann's rule is not general and temperature is not a dominant driver of biogeographic variation in mass.”

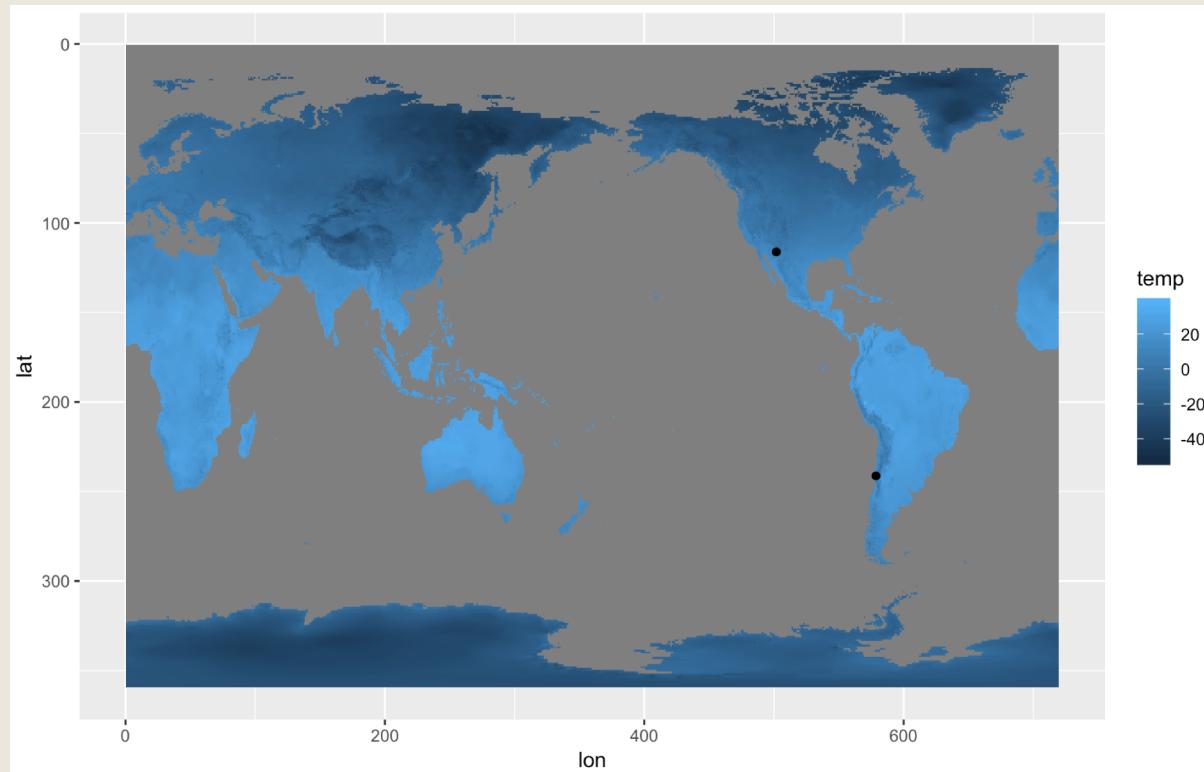


Results from data-driven approach

# Best practices in scientific computing

- Understandable
- Reproducible
- Open

# Current project

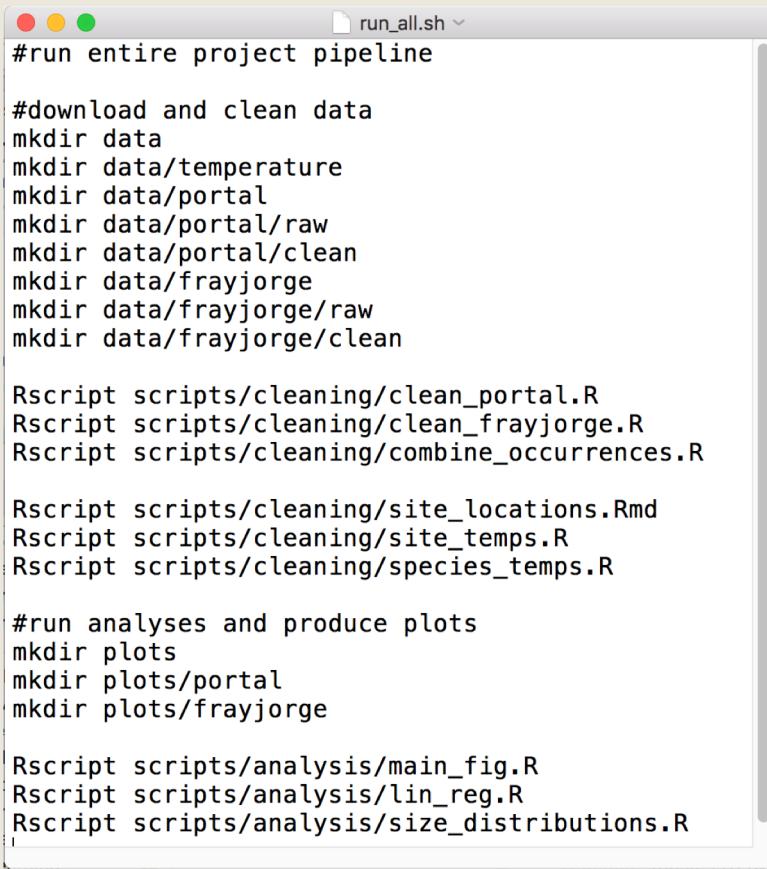


How mammals' sizes respond to temperature across time

# Making this project more understandable...

- Modular code
- Good documentation
- Metadata
- Dynamic outputs

# ...more reproducible...



```
#run entire project pipeline

#download and clean data
mkdir data
mkdir data/temperature
mkdir data/portal
mkdir data/portal/raw
mkdir data/portal/clean
mkdir data/frayjorge
mkdir data/frayjorge/raw
mkdir data/frayjorge/clean

Rscript scripts/cleaning/clean_portal.R
Rscript scripts/cleaning/clean_frayjorge.R
Rscript scripts/cleaning/combine_occurrences.R

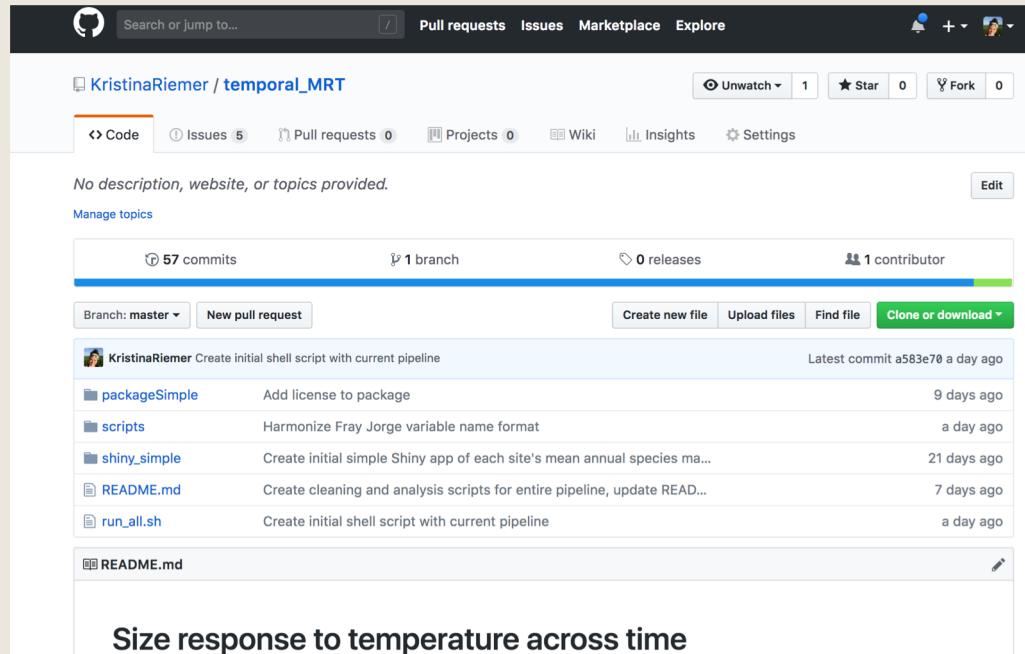
Rscript scripts/cleaning/site_locations.Rmd
Rscript scripts/cleaning/site_temps.R
Rscript scripts/cleaning/species_temps.R

#run analyses and produce plots
mkdir plots
mkdir plots/portal
mkdir plots/frayjorge

Rscript scripts/analysis/main_fig.R
Rscript scripts/analysis/lin_reg.R
Rscript scripts/analysis/size_distributions.R
```

- Clear pipeline of inputs and outputs
- Complete description of this

# ...and more open!



- Publicly available data
- Real time tracking in public GitHub repo

# So what's next?

- Testing and continuous integration
- Literate programming for manuscripts
- Documenting environments

# Scientific software development

Understandable +  
reproducible + open =  
 **reusable**

# Motivation to train scientists

- Immense benefits for my own work
- Progress not perfection
- Researchers need these skills from good instructors and good materials

## (More) motivation to train scientists

- Years being peer writing tutor
- Recently acquiring deep and broad computational skillset
- Really know it if I can teach it

# Intro to computational skills course

The screenshot shows the homepage of the Data Carpentry for Biologists website. The header features the title "Data Carpentry for Biologists" in white text on a dark background. Below the title, a subheader reads "Teaching the tools to get computers to do cool science". A navigation menu includes links for "Getting Started", "Course Materials", "Schedule", "About / Contact Us", and "In-class Feedback". A Creative Commons BY license logo is at the bottom.

**Data Carpentry for Biologists**

Teaching the tools to get computers to do cool science

- ▶ Getting Started
- 📄 Course Materials
- 📅 Schedule
- 👤 About / Contact Us
- 👉 In-class Feedback

(CC BY)

This website hosts introductory material for teaching biologists how to interact with data including: data structure, database management systems, and programming for data manipulation, analysis, and visualization. It is designed to be used as a flipped university course and also to be useful for self-guided students. Instructors are welcome to modify and use the material for your own courses. We encourage collaborative development and contributions by instructors, with the hopes that this will lead to better training and resources for everyone.

#### For Students

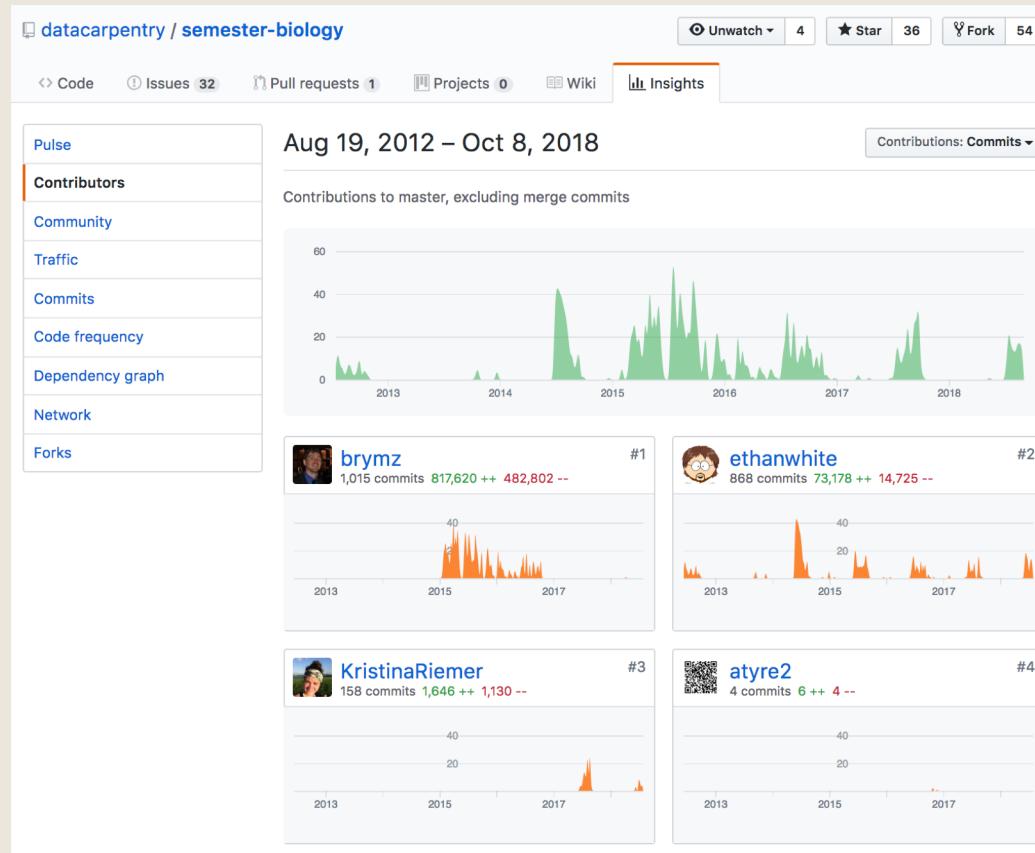
- [Syllabus](#)
- [Schedule](#)
- [Setup](#)
- [Assignment Submission & Checklist](#)
- [Self-Guided Start-Up Guide](#)
- [Datasets](#)

#### For Instructors

- [Readings](#)

- Instructor in fall 2017
- Databases, git, programming, project management, spatial data
- Really positive feedback from students

# Open-source curriculum development



# Additional curriculum development

- Initial hackathons for reproducible research workshop materials and geospatial workshop materials
- Official workshops for The Carpentries
  - See lessons on [datacarpentry.org](http://datacarpentry.org)
- Teaching newest release of geospatial materials

# What are The Carpentries?

- 2-day workshops
- Volunteer run



THE  
CARPENTRIES

# Community building

- UF Carpentries Club
- Volunteer grad students and postdocs
- Fundraising, stakeholder needs, travel grants



Home Moments Notifications 2 Messages  Search Twitter  Tweet



Tweets 187 Following 68 Followers 93 Likes 20 Lists 0 Moments 0 [Edit profile](#)

**UF Carpentries** @UFCarpentries  
Official account for the @thecarpentries community at @UF  
Gainesville, FL uf-carpentry.org Joined June 2018

**Tweets** **Tweets & replies** **Media**

You Retweeted **Hao Ye uses a flip phone** 📱 @Hao\_and\_Y · 23h  
The UF Open Source Club is hosting a in-person event fro #Hacktoberfest on October 20, from 10am - 6pm!  
[facebook.com/events/6451837...](http://facebook.com/events/6451837...)

Follow 

**Who to follow** · Refresh · View all

Justin J Millar @justinjmillar Follow 

Joan Meiners @be... Follow 

Vanessa Hull @hull\_wildlife Follow 

# Communities enable good collaboration

- Lots of collaborative work
- Range of experience levels
- Remote work

The screenshot shows a GitHub interface with the 'Issues' tab selected, displaying 8 open pull requests. The pull requests are listed with their titles, descriptions, and creation dates. A modal window titled 'Project boards for your issues and pull requests' is overlaid on the page, encouraging users to manage their workflow with project boards.

Issue Title	Description	Opened By	Comments
Permutation test results	#20 opened on Jun 19 by KristinaRiemer	KristinaRiemer	5
make mclust not give fitting status bar	#19 opened on Dec 13, 2017 by sdtaylor	sdtaylor	4
Pick ecoregion	#17 opened on Nov 9, 2017 by KristinaRiemer	KristinaRiemer	4
Per species phenology	#11 opened on Sep 26, 2017 by sdtaylor	sdtaylor	8
Trait distribution across ecoregions	#10 opened on Sep 21, 2017 by KristinaRiemer	KristinaRiemer	5
Rolling meeting to-do lists	#7 opened on Apr 20, 2017 by KristinaRiemer	KristinaRiemer	10

## Teaching computational thinking & skills



Doing understandable,  
reproducible, & open projects

Collaborating on  
computational  
projects

Developing &  
improving scientific  
software

# Thanks for coming to my talk!

Feel free to contact me:

 @KristinaRiemer

 @KristinaRiemer

kristina.riemer@weecology.org

## Acknowledgements

- Dr. Ethan White, PhD advisor
- Weecology lab members, including Dr. Morgan Ernest, Dr. Erica Christensen, and Joan Meiners
- UF Carpentries Club board members, past and present
- Special thanks to Dr. David LeBauer for the invitation to be here today

